Bioinformatics Advance Access published January 31, 2007

Databases and Ontologies

Ontology development for biological systems: Immunology

Alexander D. Diehl[*1], Jamie A. Lee[2], Richard H. Scheuermann[2,3], and Judith A. Blake[1]

[1]Mouse Genome Informatics, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04605, USA, [2]Department of Pathology, [3]Division of Biomedical Informatics, U.T. Southwestern Medical Center, 5323 Harry Hines Blvd. Dallas, TX 75390-9072, USA.

Associate Editor: John Quackenbush

ABSTRACT

Summary: We recently implemented improvements to the representation of immunology content of the biological process branch of the Gene Ontology (GO). The aims of the revision were to provide a comprehensive representation of immunological processes and to improve the organization of immunology related terms in the GO to match current concepts in the field of immunology. With these improvements, the GO will better reflect current understanding in the field of immunology and thus prove to be a more valuable resource for knowledge representation in gene annotation and analysis in the areas of immunology related to genomics and bioinformatics.

Availability: http://www.geneontology.org

Contact: adiehl@informatics.jax.org

1 INTRODUCTION

The Gene Ontology (GO) is a controlled vocabulary of terms widely used for annotation of attributes of genes and gene products across all species in biology (Gene Ontology Consortium, 2000, 2006). The GO is composed of three ontologies: 1) the cellular component ontology, which can be used to describe gene products according to their cellular localization or as part of a protein complex, 2) the molecular function ontology, which includes terms describing enzymatic and binding activities, and 3) the biological process ontology, which contains terms describing any series of events in a cell or organism accomplished by one or more ordered assemblies of molecular functions. These three domains of the GO each consist of terms organized in a hierarchy of subsumption (is_a) or parthood (part_of) relationships. The overall structure is that of a directed acyclic graph and allows for reasoning based on the relationships between the terms. The assignment of a GO term to a gene product is based either on knowledge of that gene product gained through direct experimentation as reported in the literature, or through sequence similarity to another gene product that itself has been experimentally described.

GO-based gene annotation has proved extremely useful in the systems-level interpretation of high dimensional data sets. For example, co-clustering of GO terms in gene clusters identified in gene expression microarray experiments can provide a link between the experiment conditional variables (e.g. ligand treatment) and the underlying biological process responses (Lee et al., 2006). GO term enrichment has also been useful in identifying functional modules within protein-protein networks (Luo et al., 2006).

GO terms describing processes, functions, and cellular components related to the immune system have existed in the GO from the beginning of its development, and have been used extensively in the annotation of gene products. However, particularly in the biological process ontology, the initial set of terms relating to immunology failed to cover the breadth of known immunological processes, and in many cases diverged from current usage and understanding in the field, as these terms were largely created by non-immunologists using older references. Thus, we undertook a major revision of the representation of immunology within the biological process ontology of the GO. This work developed as a joint effort of the Gene Ontology Consortium and the ImmPort project (www.immport.org) with the goal of facilitating the use of the GO in the annotation of immunological data across all species.

2 METHODS

In preparing the revision to the immunology terms in the GO, both "top-down" and "bottom-up" approaches were used. In the top-down approach, authoritative textbooks (Paul, 2003, Janeway et al., 2005) and many current reviews of specific subject areas within immunology were consulted as a basis for providing a set of high level terms in the biological process ontology to cover fundamental large scale processes that occur in the functioning of the immune system and its cells. The existing terms concerned with immunological processes in the GO were reviewed in the context of the published literature, and certain terms were either renamed or redefined to match their usage in the literature more precisely. Many new terms were introduced to match processes described within the immunology corpus.

The bottom up approach involved identifying missing terms while doing annotation of gene products, and adding these terms to the GO biological process ontology; this is otherwise known as "annotation driven ontology development." In this manner, we collected a series of possible terms related to immunology over a period of two years that were needed for annotation but did not fit into the existing structure of the biological process ontology. Most of these terms concerned lower-level processes of the immune system. In combination with the top-down approach to improving the high level structure of terms describing immunological processes, we were able to add these more granular terms in a systematic way.

After a preliminary revised structure was prepared for the set of immunology terms in the biological process ontology, the proposal was discussed at the GO Content Meeting of November 2005 held at The Institute for Genomic Research (TIGR) with a number of invited external immunologists and ontology experts from the Gene Ontology Consortium. At this meeting, the basic high level structure was finalized, and definitions for key terms were agreed upon.

Following this GO Content Meeting, additional terms were created to fill out the structure and provide for more complete coverage of immunological processes. We also wrote definitions for all new terms, supported by one or more references in the literature. Also we provided process regulation terms per GO Consortium guidelines (http://www.geneontology.org) for many of the basic process terms where we thought such terms would be of immediate utility for annotation; such regulation terms allow for more

*To whom correspondence should be addressed.

accurate annotation of the role of gene products, such as cytokines, that regulate particular processes, such as isotype-switching, but do not participate directly in the molecular mechanics of the processes. Additional rounds of review of the new terms occurred using the Gene Ontology specific pages of the SourceForge system (www.sourceforge.org), which allowed for community discussion and download of the finalized revision. In September 2006 the revised terms were incorporated into the official GO and became available for all users of the Gene Ontology.

3 RESULTS AND DISCUSSION

We have implemented a major revision to the GO biological process ontology to improve the representation of immunological processes. The revision includes 726 new GO biological process terms covering immunological processes in animals and plants, as well as the incorporation of large scale rearrangements and revisions of existing terms that rationalize term hierarchies and match term usage with current community usage for describing immunological processes. The revision also includes changes to GO biological process terms covering 'response to' and 'detection' of various organisms by other organisms. These terms are often used in annotation in conjunction with GO terms related to immunology.

A new high level term in the biological process ontology, "immune system process," has been created to group all processes directly related to the functioning of the immune system, including developmental and tolerance processes in addition to the activation and effector mechanisms of the immune system. Terms covering antigen sampling, processing, and presentation, which operate continuously, also fall under this high-level term, as well as basic terms such as "leukocyte migration" and "leukocyte homeostasis," which cover processes that occur both apart from and as part of immune responses.

The term "immune response," which previously existed in the GO, is now a child of the "immune system process" term. Children of the "immune response" term include the basic processes of "innate immune response," "adaptive immune response," and "humoral immune response," as well as a grouping term for organ or tissue specific immune responses such as those in the mucosa. Because the GO allows for multiple inheritance, many subterms are children of more than one of these terms covering basic types of immune response. An example is the GO term "humoral immune response mediated by circulating immunoglobulin," which has is_a relationships to both "humoral immune response," as a direct child, and "adaptive immune response," via several intermediate terms. Such dual parentage accurately reflects how these processes are alternatively regarded by working immunologists.

Activation of the immune response, in particular cell surface-linked signaling pathways involved in leukocyte activation is covered by a new hierarchy of terms. Similarly, terms for immune effector mechanisms have been collected under the "immune effector process" parent term. These terms cover both common mechanisms employed in the execution of an immune response, such as degranulation, as well as cell-type specific mechanisms, such immunoglobulin production, which is specific for B lymphocytes.

The new structure contains a greatly expanded number of terms covering subprocesses of the innate immune response, including processes for activation mechanisms such as TLR signaling pathways, which were not previously covered in the GO. Also, existing GO terms that refer to plant innate immune processes have been collected under the general "innate immune response" hierarchy, reflecting current theory in the plant field (Nurnberger et al., 2004). Similarly, existing GO terms that refer to specific invertebrate innate immune process are now collected here as well.

Expanded hierarchies cover developmental processes of the immune system including a greatly increased number of terms describing B and T cell differentiation. Tolerance induction, both central and peripheral, is now fully represented. In addition, many terms have been added to cover somatic diversification of immune receptors processes, both for immunoglobulins and for the different systems of immune receptors found in lampreys and invertebrates (Pancer, 2006).

Terms were also added to the "inflammatory response" subtree that more fully describe both processes that occur during acute and chronic inflammatory responses, and processes that differentiate between an inflammatory response initiated by antigen vs. one initiated by a non-antigenic stimulus.

The revised structure will, importantly, allow addition of new terms in a consistent fashion. For example, new terms related to surface signal transduction for immune receptors or terms related to T cell mediated immunity or cytokine production have an obvious placement in the revised ontology. Also, further refinement of regulation terms is currently underway as part of a larger ontological analysis and restructuring of the GO in all areas.

The revisions provide many benefits for annotators who use the GO for annotation of gene products in that many new terms are available and all the terms related to immunology are now named and organized in a more logical fashion that parallels usage in the immunological literature. As the new terms are increasingly used for annotation, interpretive analysis of high-throughput experiments (e.g. gene expression microarrays) in the area of immunology will be improved as more granular annotation will be available for many gene products. Furthermore, these improvements to the relational hierarchy for immunology-related GO terms should also improve the performance of data mining algorithms.

ACKNOWLEDGEMENTS

We would like to thank the participants in the GO Content Meeting, November 15-16, 2005, at TIGR (an attendee list can be found at http://www.geneontology.org/GO.meetings.shtml), the Immune Epitope Database, the GO Consortium, and the BISC development team. This work was supported by NIH grants HG002273 (A.D.D. and J.A.B.) and N01AI40076 (J.A.L. and R.H.S.).

REFERENCES

Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nature Genet.,* 25, 25-29.

Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.,* 34, D322-D326.

Janeway, C.A., Travers, P., Walport, M., Shlomchik, M.J. (2004) *Immunobiology,* Garland Science, New York.

Lee, J.A. et al. (2006) Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. *BMC Bioinformatics,* 7, 237.

Luo, F. et al. (2006) Modular decomposition of protein interaction networks. *Bioinformatics*, doi: 10.1093/bioinformatics/btl562.

Nurnberger, T. et al. (2004) Innate immunity in plants and animals: striking similarities and obvious differences. *Immunol. Rev.,* 198, 249-66.

Paul, W.E. (ed.) (2003) *Fundamental Immunology,* Lippencott Williams & Wilkins, NewYork.

Pancer Z., Cooper M.D. (2006) The evolution of adaptive immunity. *Annu. Rev. Immunol.,* 24, 497-518.